RUC at TREC 2014: Select Resources Using Topic Models

Qiuyue Wang, Shaochen Shi, Wei Cao
School of Information
Renmin University of China
Beijing 100872
{qiuyuew, shishaochen, caowei}@ruc.edu.cn

ABSTRACT

This paper describes the work done in Renmin University of China for the Federated Web Search Track of TREC 2014. We participated in the resource selection task. We used the LDA topic modeling approach to select potentially relevant resources for a given query. The initial results are promising.

1. INTRODUCTION

The Federated Web Search Track [1] is intended to investigate federated search techniques in a realistic Web setting with a large number of online Web search services. This year the track contains three tasks: resource selection (select relevant resources for a given query), vertical selection (classify each query into a fixed set of 24 verticals) and results merging (merge results returned by several resources into a single ranked list). We participated in the resource selection task.

The input to the resource selection task is a list of 149 online search engines and a collection provided by the organizers consisting of sampled query results (pages and snippets) obtained by sampling 4000 queries on each of the 149 resources. Given a test query, the task is to return a list of resources ranked by their capabilities of returning relevant results for the query.

The potential relevance of a resource to a given query can be estimated based on many factors [9], for example the authority or usefulness of the resource (which is query independent), its content relevance to the query based on text matching or on topical matching with the query.

Most previous approaches rank the resources according to their content relevance to the query mainly based on text matching, e.g. big-document approaches like CORI [2] and [4], and small-document approaches like GAVG [8], ReDDE [5] and CRCS [6]. In big-document approaches, each resource is represented as a single large document by concatenating all its sampled documents. Thus the resources can be ranked using any existing document retrieval models. In small-document approaches, the sampled documents for each resource are not concatenated but scored individually. Each resource is then ranked based on the matching scores of its sampled documents, either by aggregating the scores or by estimating the density of relevant documents based on the scores.

All these methods are based on text matching between sampled documents and query, which suffer from the problem of missing vocabulary in the relatively small samples for each resource. Some other approaches address this problem by describing each resource by the categories or topics that it covers and ranking the resources based on topical matching between the query and resource. By modelling resources in a low dimensional topic space, the model can generalize well to unseen documents and thus alleviate the problem of incomplete information caused by small samples.

Many existing category or topic based approaches make use of predefined category hierarchies [7][11], e.g. Open Directory Project (ODP) or KDD-CUP 2005. For instance, [11] uses the online service of ODP to get the list of ODP categories for each resource and query, and ranks the resource by the similarity of its category vector to that of the query using cosine or Jaccard similarities. In [7], resources are first classified into a predefined topical hierarchy by focused probing. Statistical summary for each resource is then smoothed with a set of topically related resources in the topical hierarchy to alleviate the problem of sparse samples.

Some other topic based approaches use the extracted topics from the data, e.g. by clustering the documents [3] or by topic modelling approach [10]. In [10], the authors introduce a hierarchical extension to LDA, which models the generative process of resources explicitly. However, the trained topic models are applied to smooth the estimated term distribution in text matching instead to match the topic distribution directly.

At TREC 2014, we used the topic modeling approach based on LDA to return the ranked list of resources. First, we train a LDA-based topic model over the collection of sampled documents for all the resources, and obtain the topic distribution for each resource. Given a query, we expand the query using Google Search API, and then infer the topic distribution of the expanded query based on the previously trained topic model. Finally, the resources are ranked according to the distance of their topic distributions to that of the query.

2. Resource Selection

2.1 Topic Models

There are various probabilistic topic models used in many applications. Latent Dirichlet Allocation (LDA) is the most frequently used topic model. It is a probabilistic generative model for documents: each document is modelled as a mixture of topics where each topic is a probability distribution of words. The LDA model specifies a probabilistic procedure by which documents can be generated, thus it has the ability to generalize well to unseen documents. Given a set of observed documents, statistical techniques can be applied to infer the latent topic model that is most likely to have generated the observed data. This involves inferring the word distribution for each topic and topic distribution for each document.

2.2 Resource Representation

For the problem of resource selection, we are interested in knowing the latent topic structure for each resource. Each resource consists of a collection of documents with a small number of them being observed (i.e. sampled). To infer the topic

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
RUC at TREC 2014: Select Resources Using Topic Models	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Renmin University of China,School of Information,Beijing 100872,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
presented in the proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014) held in Gaithersburg, Maryland, November 19-21, 2014. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

14. ABSTRACT
This paper describes the work done in Renmin University of China for the Federated Web Search Track of TREC 2014. We participated in the resource selection task. We used the LDA topic modeling approach to select potentially relevant resources for a given query. The initial results are promising.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**3**	

distribution of resources, one can model the generation of resources explicitly in a multi-level topic model like the MCTM in [10]. Alternatively in this paper, we investigated two simpler strategies to infer the topic distribution for each resource: big-document and small-document strategies.

In the big-document strategy, all the sampled documents for each resource are concatenated to form a single large document. Then the topic model for generating the 149 large documents, each representing a resource, are inferred using a standard Gibbs sampling algorithm implemented in MALLET [13].

At TREC FedWeb 2014, each resource is sampled by 4000 queries with top 10 pages returned for each query being collected, the size of the resulting large document for each resource could be of hundreds of megabytes or even several gigabytes, which renders the inference of topic models on these large documents very inefficient. Inspired by some previous study on predicting page relevance by its snippet [14][15], we concatenate all the top 10 snippets instead of pages returned for each sample query to form the large document for a resource. There are 40,000 sampled snippets for each resource at TREC 2014, but the sizes of the snippets are far smaller than the corresponding Web pages. In addition, for some search engines, like the resource *e122* (Picasa) in FedWeb 2014, all the sampled pages are non-text files, e.g. image or video files, so the big-documents for such engines by concatenating the text from all its sampled pages would be empty, which causes such resources would not be selected for any queries. When we concatenate the sampled snippets however, the titles and descriptions of the snippets would offer textual depictions about the sampled data so that the resource could still have the chance of being selected for relevant queries. The effectiveness of using snippets to represent resources is demonstrated in the experiment results in Section 3. We name this strategy as snippet-based big-document strategy.

In the small-document strategy, we treat all the sampled documents for all the resources as one collection and train the topic model over the collection to get the topic distribution for each sampled document. Then we represent the topic distribution of each resource using the mean topic distributions of all its sampled documents. We did not investigate the snippet-based small-document strategy because the snippets are short and LDA topic models in general suffer from the sparse word co-occurrence in short texts.

Before applying MALLET to train topic models, either on the set of large documents or small documents, we preprocess the data by parsing the pages (html, txt, doc, xls, ppt, pdf, xml files) into tokens, removing the stopwords listed in the Indri's standard stopword list, and stemming the tokens with the Krovetz stemmer.

2.3 Query Representation

Given a test query, we infer a topic distribution for it using the topic model trained on the document collection as described in Section 2.2. As the query is typically very short, it is hard to infer its topic distribution accurately. To overcome the problem, we first expand the given query using Google Search API. We submit the query to the Google Search API, and collect the top 10 snippets returned by the API. The top 50 most frequent terms occurring in the 10 snippets are selected to expand the query. We preprocess the expanded query using the same tokenization, stopword removal and stemming as that used for document preprocessing, and infer the topic distribution for the query with MALLET.

2.4 Resource Ranking and Selection

With the topic distributions inferred for a resource and query, the relevance of the resource to the query can be measured by the extent that they share the same topics, i.e. the similarity between their topic distributions. The rationale behind this is that if a resource is more topically similar to the query, it is more likely to return relevant results for the query.

A standard function to measure the difference between two probability distributions is the Kullback Leibler (KL) divergence. We compute the KL divergence between the topic distributions of resource R and query Q as the following and rank the resources accordingly.

$$D_{KL}(Q \mid\mid R) = \sum_{i=1}^{K} P(t_i \mid Q) \frac{P(t_i \mid Q)}{P(t_i \mid R)} \quad (1)$$

where K is the number of topics, which is an input parameter to the LDA topic model. In Section 3, we evaluate the performance with different K values.

3. Results

Table 1 shows the results obtained by evaluating our resource selection approaches on the FedWeb 2013 collection. We evaluate the three strategies of generating resource representations as discussed in Section 2.2, with varying numbers of topics (K) in training the LDA topic model. The "engines" column shows the results of the runs generated using the big-document strategy; "search" column is about all the runs generated by the snippet-based big-document strategy; and "docs" column presents the results of the runs generated by the small-document strategy. The performance of runs is measured by the nDCG@20, which is the main evaluation metric used at the FedWeb research selection task.

Table 1. Performance of variations of the approach on the FedWeb 2013 collection

number of topics (K)	nDCG@20		
	engines	**search**	**docs**
10	0.263	0.245	0.110
20	0.282	0.203	0.163
30	0.337	0.296	0.221
50	-	0.314	0.244
75	-	0.333	0.263
100	-	**0.372**	0.260
125	-	0.369	-
150	-	0.366	-
200	-	0.307	-
250	-	0.328	-

With the increasing number of topics, i.e. larger K, training LDA topic models for the big- and small-document strategies becomes more and more inefficient. We failed in generating the results in some cases within a reasonable amount of time. The missing

results are shown as "-" in Table 1. We can observe that big-document strategy is the most effective one for resource selection, but also the most expensive one. Alternatively snippet-based big-document strategy is much cheaper with only a slight degradation on performance.

We submitted 6 runs to the TREC FedWeb 2014. Their official evaluation results are shown in Table 2. Three "FW14Search*" runs are generated using the snippet-based big-document strategy, while the other three "FW14Docs*" runs generated by the small-document strategy. The different numbers in the run IDs are the different numbers of topics (*K*) set in training topic models. We did not submit any big-document runs because the average length of the big-documents for resources in FedWeb 2014 almost doubles that in FedWeb 2013, which renders it a much less feasible strategy.

Similar to the results on the FedWeb 2013 collection, snippet-based big-document strategy is more effective than the small-document strategy. The optimal number of topics in FedWeb 2014 is different from that in 2013 however. The performance in terms of nDCG@20 achieves the best when the number of topics is 50 in FedWeb 2014, while the optimal number of topics is 100 on the collection of FedWeb 2013. Thus the problem of how to choose the right number of topics arises, which we leave for future work.

Our runs are the second best group of runs at FedWeb 2014, right after the runs submitted by East China Normal University. The good performance of their runs largely depends on a query-independent prior ranking of the resources learned on the results from FedWeb 2013. Such query-independent factors are orthogonal to our approach, so combination of the two could probably further improve the performance.

Table 2. Performance of variations of the approach on the FedWeb 2014 collection

Run ID	nDCG @20	nDCG @10	nP@1	nP@5
FW14Search100	0.505	0.425	**0.278**	0.384
FW14Search75	0.461	0.366	0.256	0.345
FW14Search50	**0.517**	**0.426**	0.271	**0.404**
FW14Docs100	0.444	0.337	0.165	0.239
FW14Docs75	0.422	0.306	0.106	0.198
FW14Docs50	0.419	0.292	0.174	0.203

4. Conclusion and Future work

In this paper, we describe our participation in the TREC 2014 Federated Web Search Track. We propose to use latent topic modelling approaches, e.g. LDA, to discover the topic structure for each resource and rank the resources with respect to a query based on their topic distribution similarities. To be able to infer topic distributions for very short queries, we expand the query using Google Search API. The initial results are promising. In contrast to the common findings in most text-matching methods, big-document strategy for resource representation is more effective the small-document strategy in our topic-matching approach. More encouraging finding is that snippet-only representation of resources can achieve very good performance, which makes this approach more feasible to be used in practice.

As for our future work, we would like to further investigate various other topic modelling approaches, e.g. the hierarchical LDA like the MCTM proposed in [10], ESA and so on. Combining text-matching, topic-matching, and prior ranking of resources offers another interesting opportunity for future work.

5. ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (No. 61202331).

6. REFERENCES

[1] https://sites.google.com/site/trecfedweb/ .

[2] J. P. Callan, Z. Lu, and W. B. Croft. Searching Distributed Collections With Inference Networks. In SIGIR 1995, pages 21-28.

[3] J. Xu and W. B. Croft, Cluster-Based Language Models for Distributed Retrieval, In SIGIR 1999.

[4] L. Si, R. Jin, J. Callan, and P. Ogilvie. A Language Modeling Framework for Resource Selection and Results Merging. In CIKM 2002, pages 391-397.

[5] L. Si and J. Callan. Relevant Document Distribution Estimation Method for Resource Selection. In SIGIR 2003, pages 298-305.

[6] M. Shokouhi. Central-Rank-Based Collection Selection in Uncooperative Distributed Information Retrieval. In ECIR 2007, pages 160-172.

[7] P. G. Ipeirotis and L. Gravano. Classification-Aware Hidden-Web Text Database Selection. ACM Trans. Inf. Syst. Vol. 26, No. 2, Article 6, April 2008.

[8] J. Seo and B. W. Croft. Blog Site Search Using Resource Selection. In CIKM 2008, pages 1053-1062.

[9] J. Arguello, J. Callan, and F. Diaz. Classification-Based Resource Selection. In CIKM 2009, pages 1277-1286.

[10] M. Baillie, M. Carmen, and F. Crestani. A Multiple-Collection Latent Topic Model for Federated Search. Information Retrieval (2011) 14:390-412.

[11] A. Bellogin, G. G. Gebremeskel, J. He, A. Said, T. Samar, A. P. de Vries. CWI and TU Delft at TREC 2013: Contextual Suggestion, Federated Web Search, KBA, and Web Tracks. In TREC 2013.

[12] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research 3 (4–5): pp. 993–1022.

[13] A. K. McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. 2002.

[14] T. Demeester, D. Nguyen, D. Trieschnigg, and D. Hiemstra. What Snippets Say About Pages in Federated Web Search. In AIRS 2012.

[15] T. Demeester, D. Nguyen, D. Trieschnigg, C. Develder and D. Hiemstra. Snippet-based Relevance Predictions for Federated Web Search. In ECIR 2013.